Information and Inference: A Journal of the IMA (2012) Page 1 of 19
doi:10.1093/imaiai/iarxxx

The Masked Sample Covariance Estimator: An Analysis via Matrix Concentration Inequalities

RICHARD Y. CHEN,
ycchen@caltech.edu

ALEX GITTENS
gittens@cms.caltech.edu

AND

JOEL A. TROPP
Corresponding author: jtropp@cms.caltech.edu

Dept. of Computing and Mathematical Sciences
California Institute of Technology
1200 E. California Blvd., MC 305-16
Pasadena, CA 91125-5000, USA

[Received on XX XX XXXX; revised on XX XX XXXX; accepted on XX XX XXXX]

Covariance estimation becomes challenging in the regime where the number p of variables outstrips the number n of samples available to construct the estimate. One way to circumvent this problem is to assume that the covariance matrix is nearly sparse and to focus on estimating only the significant entries. To analyze this approach, Levina and Vershynin (2011) introduce a formalism called *masked covariance estimation*, where each entry of the sample covariance estimator is reweighted to reflect an *a priori* assessment of its importance.

This paper provides a short analysis of the masked sample covariance estimator by means of a matrix concentration inequality. The main result applies to general distributions with at least four moments. Specialized to the case of a Gaussian distribution, the theory offers qualitative improvements over earlier work. For example, the new results show that $n = \mathrm{O}(B \log^2 p)$ samples suffice to estimate a banded covariance matrix with bandwidth B up to a relative spectral-norm error, in contrast to the sample complexity $n = \mathrm{O}(B \log^5 p)$ obtained by Levina and Vershynin.

Keywords: Covariance estimation, matrix concentration inequality, matrix Khintchine inequality, matrix Rosenthal inequality, random matrix, Schur product.

2010 Math Subject Classification. Primary: 60B20; Secondary: 62H12, 60F10, 60G50.

1. Introduction

A fundamental problem in multivariate statistics is to obtain an accurate estimate of the covariance matrix of a multivariate distribution given independent samples from the distribution. This challenge arises whenever we need to understand the spread of the data and its marginals, for example, when we perform regression analysis [12] or principal component analysis [18].

In the classical setting where the number of samples exceeds the number of variables, the behavior of standard covariance estimators is well understood [17, 28, 29]. The random matrix literature also

contains a substantial amount of relevant work; we refer to the book [2] and the survey [39] for further information.

Modern applications, in contrast, often involve a small number of samples and a large number of variables. The paucity of data makes it impossible to obtain an accurate estimate of a general covariance matrix. As a remedy, we must frame additional model assumptions and develop estimators that exploit this extra structure. Over the last few years, a number of papers, including [4, 5, 8, 11, 13, 33], have focused on the situation where the covariance matrix is sparse or nearly so. In this case, we imagine that we could limit our attention to the significant entries of the covariance matrix and thereby perform more accurate estimation with fewer samples.

This paper studies a particular technique for the sparse covariance problem that we call the *masked sample covariance estimator*. This approach uses a mask matrix, constructed *a priori*, to specify the importance we place on each entry of the covariance matrix. By reweighting the sample covariance estimate using a mask, we can reduce the error that arises from imprecise estimates of covariances that are small or zero. The mask matrix formalism was introduced by Levina and Vershynin [24] to provide a unified treatment of some earlier methods for sparse covariance estimation; we refer to their paper for a more detailed discussion of prior work.

This paper provides a new analysis of the masked sample covariance estimator using some recent ideas from random matrix theory [1, 22, 27, 30, 34, 35, 37–39]. These methods, collectively known as *matrix concentration inequalities*, are particularly well suited for studying a sum of independent random matrices. The results provide strong bounds on the moments and exponential moments of the spectral norm of the sum by harnessing information about the individual summands. Indeed, matrix concentration inequalities can be viewed as far-reaching extensions of the classical inequalities for a sum of scalar random variables [10]. As we demonstrate in this work, matrix concentration inequalities sometimes allow us to replace devilishly hard calculations with simple arithmetic.

One of our main reasons for writing this paper is to show that matrix concentration inequalities can streamline the analysis of random matrices that arise in statistical applications. We believe that the simplicity of our arguments and the strength of our conclusions make a compelling case for the value of these methods. Indeed, we hope that matrix concentration inequalities will find a place in the toolkit of researchers working on multivariate problems in statistics.

1.1 *Classical Covariance Estimation*

Consider a random vector

$$\boldsymbol{x} = (X_1, X_2, \dots, X_p)^* \in \mathbb{R}^p.$$

Let $\boldsymbol{x}_1, \dots \boldsymbol{x}_n$ be independent random vectors that follow the same distribution as $\boldsymbol{x}$. For simplicity, we assume that the distribution is known to have zero mean: $\mathbb{E}\,\boldsymbol{x} = \mathbf{0}$. The *covariance matrix* $\boldsymbol{\Sigma}$ is the $p \times p$ matrix that tabulates the second-order statistics of the distribution:

$$\boldsymbol{\Sigma} := \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^*), \tag{1.1}$$

where * denotes the transpose operation. The classical estimator for the covariance matrix is the *sample covariance matrix*:

$$\widehat{\boldsymbol{\Sigma}}_n := \frac{1}{n} \sum\nolimits_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^*. \tag{1.2}$$

The sample covariance matrix is an unbiased estimator of the covariance matrix: $\mathbb{E}\,\widehat{\boldsymbol{\Sigma}}_n = \boldsymbol{\Sigma}$.

Given a tolerance $\varepsilon \in (0,1)$, we can study how many samples n are typically required to provide an estimate with relative error ε in the spectral norm:

$$\mathbb{E} \left\| \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma} \right\| \leqslant \varepsilon \left\| \boldsymbol{\Sigma} \right\|. \tag{1.3}$$

This type of spectral-norm error bound is quite powerful. It limits the magnitude of the estimation error for each entry of the covariance matrix; it provides information about the variance of each marginal of the distribution of $\boldsymbol{x}$; it even controls the error in estimating the eigenvalues of the covariance using the eigenvalues of the sample covariance.

Unfortunately, the error bound (1.3) for the sample covariance estimator demands a lot of samples. Indeed, suppose that the covariance matrix has full rank. When $n < p$, the sample covariance is rank-deficient, so the spectral norm error is bounded away from zero!

Typical positive results state that the sample covariance estimator is precise when the number of samples is proportional to the number of variables, provided that the distribution decays fast enough. For example, assuming that $\boldsymbol{x}$ follows a normal distribution,

$$n \geqslant \mathrm{C}\, \varepsilon^{-2} p \quad \Longrightarrow \quad \left\| \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma} \right\| \leqslant \varepsilon \left\| \boldsymbol{\Sigma} \right\| \quad \text{with high probability.} \tag{1.4}$$

We use the convention that C denotes an absolute constant whose value may change from appearance to appearance. See [39, Thm. 57 et seq.] for details of obtaining the bound (1.4).

1.2 *The Masked Sample Covariance Estimator*

In the regime $n \ll p$, where we have very few samples, we cannot hope to achieve an estimate like (1.3) for a general covariance matrix. Instead, we must instate additional assumptions and incorporate this prior information to construct a regularized estimator. Over the last few years, researchers have studied the case where the covariance matrix is sparse or nearly sparse. In this setting, we can often refine our estimation procedure by focusing on the most significant entries of the covariance matrix.

One way to formalize this idea is to construct a symmetric $p \times p$ matrix $\boldsymbol{M}$ with real entries, which we call the *mask matrix*. In the simplest case, the mask matrix has 0–1 values that indicate which entries of the covariance we attend to. A unit entry $m_{ij} = 1$ means that we estimate the interaction between the ith and jth variables, while a zero entry $m_{ij} = 0$ means that we abdicate from making the estimate. More generally, we can allow the components of the mask to range over the interval $[0,1]$, in which case the size of m_{ij} is proportional to the importance of estimating the (i,j) entry of the covariance matrix.

Given a mask $\boldsymbol{M}$, we define the *masked sample covariance estimator* $\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n$, where the symbol $\odot$ denotes the componentwise (i.e., Schur or Hadamard) product. The following expression bounds the root-mean-square spectral-norm error that this estimator incurs.

$$\left[\mathbb{E} \left\| \boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma} \right\|^2 \right]^{1/2} \leqslant \underbrace{\left[\mathbb{E} \left\| \boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma} \right\|^2 \right]^{1/2}}_{\text{variance}} + \underbrace{\left\| \boldsymbol{M} \odot \boldsymbol{\Sigma} - \boldsymbol{\Sigma} \right\|}_{\text{bias}}. \tag{1.5}$$

The second term in (1.5) represents the bias in the estimate owing to the presence of the mask, while the first term measures how much the estimator fluctuates about its mean value. This bound is analogous with the classical bias–variance decomposition for the mean-squared-error (MSE) of a point estimator.

To obtain an effective estimator, we must design a mask that controls both the bias and the variance in (1.5). We cannot neglect too many components of the covariance matrix, or else the bias in the masked estimator may compromise its accuracy. At the same time, each additional component we

estimate contributes to the size of the variance term. In the case where the covariance matrix is sparse, it is natural to strike a balance between these two effects by refusing to estimate entries of the covariance that we know *a priori* to be small or zero.

Many of the regularization techniques for sparse covariance estimation studied in the literature, such as [5, 8, 13], can be described using mask matrices. These works focus on specific cases, such as banded masks and tapered masks, whereas we have followed Levina and Vershynin [24] by allowing an arbitrary symmetric mask $\boldsymbol{M}$. We refer to the papers cited in this paragraph for further background and references.

REMARK 1.1 (Adapted Masks) In statistical practice, it may be more natural to *estimate* the mask matrix $\boldsymbol{M}$ from the observed samples, rather than to construct the mask *a priori*. A number of authors, including El Karoui [11], have studied the performance of covariance estimators with an adaptive threshold. In this work, we focus on the simpler case where the mask is fixed. It presents an interesting challenge to analyze a data-dependent mask using matrix concentration inequalities.

1.2.1 *Example: The Banded Estimator of a Decaying Covariance Matrix.* Let us consider the case where the entries of the covariance matrix $\boldsymbol{\Sigma}$ decay away from the diagonal. Suppose that, for a fixed parameter $\alpha > 1$,

$$\left|(\boldsymbol{\Sigma})_{ij}\right| \leqslant |i-j+1|^{-\alpha} \quad \text{for each pair } (i,j) \text{ of indices.}$$

This type of property might hold for a random process whose correlations are localized in time. (That is, the current value of the process depends weakly on the past and the future.) Related covariance structures arise for random fields that have short spatial correlation scales.

A simple (suboptimal) approach to this covariance estimation problem is to focus on a band of entries near the diagonal. Suppose that the bandwidth $B := 2b+1$ for a nonnegative integer b. For instance, a mask with bandwidth $B = 3$ for an ensemble of $p = 5$ variables takes the form

$$\boldsymbol{M}_{\text{band}} := \begin{bmatrix} 1 & 1 & & & \\ 1 & 1 & 1 & & \\ & 1 & 1 & 1 & \\ & & 1 & 1 & 1 \\ & & & 1 & 1 \end{bmatrix}.$$

In this setting, it is straightforward to compute the bias term in (1.5). Indeed,

$$\left|(\boldsymbol{M} \odot \boldsymbol{\Sigma} - \boldsymbol{\Sigma})_{ij}\right| \leqslant \begin{cases} |i-j+1|^{-\alpha}, & |i-j| > b \\ 0, & \text{otherwise.} \end{cases}$$

Gershgorin's theorem [15, Sec. 6.1] implies that the spectral norm of a symmetric matrix is dominated by the maximum ℓ_1 norm of a column, so

$$\|\boldsymbol{M} \odot \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\| \leqslant 2\sum_{k>b} (k+1)^{-\alpha} \leqslant \frac{2}{\alpha-1}(b+1)^{1-\alpha}.$$

The second inequality follows when we compare the sum with an integral. A similar calculation shows that $\|\boldsymbol{\Sigma}\| \leqslant 1 + 2(\alpha-1)^{-1}$. Assuming the covariance matrix really does have constant spectral norm, it follows that

$$\|\boldsymbol{M} \odot \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\| \lesssim B^{1-\alpha}\,\|\boldsymbol{\Sigma}\|$$

On a relative scale, the bias decreases polynomially as we increase the bandwidth B of the mask. Note that this estimate follows from an easy application of classical matrix analysis.

We cannot complete the bound (1.5) for the estimation error without understanding the behavior of the fluctuation term. In contrast to the bias term, this analysis is challenging, and it requires an excursion into the field of random matrix theory.

1.3 *The Performance of Masked Covariance Estimation*

This paper studies the variance term in the error bound (1.5) for the masked sample covariance estimator. To perform this analysis, we must address a variety of issues: How does the structure of the mask $\boldsymbol{M}$ affect the performance of the estimator? How many samples n do we need to control the size of the fluctuation? What role does the distribution of the underlying random vector $\boldsymbol{x}$ play?

In Section 3, we use a matrix concentration inequality to obtain a bound for the variance term in (1.5) that holds for any distribution on $\boldsymbol{x}$ with four finite moments. In the Introduction, we focus on the simpler setting where the random vector follows a normal distribution. Our theory for this case highlights the factors that affect the performance of the estimator. We examine how the structure of the mask enters into the error bound, and we describe how the error decreases with the number of samples. We also discuss how the correlation among variables affects the difficulty of the covariance estimate.

Note that this work focuses on the random matrix aspects of the masked sample covariance estimator. As a consequence, we purposely avoid a detailed discussion of the statistical issues. In particular, we make no further study of the (deterministic) bias term in (1.5). Nor do we presume to make any claims about statistical practice.

1.3.1 *The Complexity of a Mask.*

The number of samples we need to control the variance in (1.5) depends on "how much" of the covariance matrix we are attempting to estimate. A "complex" mask requires us to estimate many interactions between variables, so we need a lot of samples to limit the fluctuation of the masked sample covariance estimator. In this section, think about masks that take 0–1 values to gain intuition.

Our analysis identifies two separate metrics that quantify the complexity of the mask. The matrix norm $\|\cdot\|_{1\to 2}$ returns the maximum ℓ_2 norm of a column. The first complexity measure is the *square* of the maximum column norm:

$$\|\boldsymbol{M}\|_{1\to 2}^2 := \max_j \left[\sum\nolimits_i m_{ij}^2\right].$$

Roughly, the bracket counts the number of interactions we want to estimate that involve the variable j, and the maximum computes a bound over all p variables. This metric is "local" in nature. The second complexity measure is the spectral norm $\|\boldsymbol{M}\|$ of the mask matrix, which provides a more "global" view of the complexity of the interactions that we estimate.

Some examples may illuminate how these metrics reflect the properties of the mask. First, suppose that we estimate the entire covariance matrix, so the mask is the matrix of ones:

$$\boldsymbol{M} = \text{matrix of ones} \quad\Longrightarrow\quad \|\boldsymbol{M}\|_{1\to 2}^2 = p \quad\text{and}\quad \|\boldsymbol{M}\| = p.$$

Next, consider the mask that arises from the banded estimator in Section 1.2.1:

$$\boldsymbol{M} = \text{0–1 matrix, bandwidth } B \quad\Longrightarrow\quad \|\boldsymbol{M}\|_{1\to 2}^2 \leqslant B \quad\text{and}\quad \|\boldsymbol{M}\| \leqslant B$$

because there are at most B ones in each row and column. When $B \ll p$, the banded mask asks us to estimate fewer interactions than the full mask, so we expect the covariance estimate to be easier.

REMARK 1.2 (Are the Two Metrics Really Different?) In the examples above, the two metrics take the same value, but this coincidence does not always occur. Although the spectral norm dominates the maximum ℓ_2 norm of a column, the *square* of the maximum column norm can be substantially larger or substantially smaller than the spectral norm.

REMARK 1.3 (Scaling) We can reduce the size of both complexity measures by rescaling the mask, but changing the scale of the mask may also increase the bias term in (1.5). When studying the masked sample covariance estimator in the context of a particular application, it is essential to consider both the variance and the bias terms.

1.3.2 *Masked Covariance Estimation for Multivariate Normal Distributions.* We are now prepared to present our main result for the case where the random vector $\boldsymbol{x}$ follows a normal distribution with zero mean. The statement involves the norm $\|\cdot\|_{\max}$, which returns the maximum absolute entry of a matrix.

THEOREM 1.1 (Masked Covariance Estimation for a Gaussian Distribution) Fix a $p \times p$ symmetric mask matrix $\boldsymbol{M}$, where $p \geqslant 3$. Suppose that $\boldsymbol{x}$ is a Gaussian random vector in $\mathbb{R}^p$ with mean zero. Define the covariance matrix $\boldsymbol{\Sigma}$ and the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_n$ as in (1.1) and (1.2). Then the variance of the masked sample covariance estimator satisfies

$$\begin{aligned}\left[\mathbb{E}\left\|\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}\right\|^2\right]^{1/2} \\ \leqslant \mathrm{C}\left[\left(\frac{\|\boldsymbol{\Sigma}\|_{\max}}{\|\boldsymbol{\Sigma}\|} \cdot \frac{\|\boldsymbol{M}\|_{1\to 2}^2 \log p}{n}\right)^{1/2} + \frac{\|\boldsymbol{\Sigma}\|_{\max}}{\|\boldsymbol{\Sigma}\|} \cdot \frac{\|\boldsymbol{M}\| \log p \cdot \log(np)}{n}\right] \|\boldsymbol{\Sigma}\|. \quad (1.6)\end{aligned}$$

The proof of Theorem 1.1 appears in Section 3.4. The rest of this section consists of some discussion of the result, as well as comparisons with related work.

1.3.3 *Extension to Distributions with Nonzero Mean.* In the actual practice of covariance estimation, we would center each sample empirically by subtracting the sample mean $\bar{\boldsymbol{x}} = n^{-1}\sum_{i=1}^n \boldsymbol{x}_i$. The sample covariance (1.2) is computed using the centered samples $\widetilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \bar{\boldsymbol{x}}$ instead of the original samples $\boldsymbol{x}_i$. Theorem 1.1 can be extended to cover the masked covariance estimator formed with centered samples. See [24, Rem. 4] for the details of the argument.

1.3.4 *Discussion and Interpretation.* Theorem 1.1 exposes several phenomena in the behavior of the masked sample covariance estimator. The first term on the right-hand side of (1.6) reflects the scale for moderate deviations of the estimator, and it depends on the "local" complexity of the mask. The second term on the right-hand side of (1.6) reflects the scale for large deviations of the estimator. It depends on the "global" complexity of the mask. When the sample size n is large, the first term drives the bound because the second term usually decays faster.

The ratio of the maximum entry of the covariance matrix to the spectral norm is an interesting feature of (1.6). The ratio never exceeds one, but it can be as small as p^{-1} when the covariance matrix has rank one. We interpret this factor as saying that covariance estimation is easier when the variables are highly correlated.

1.3.5 *Sample Complexity Bound.* Markov's inequality can be used to convert (1.6) into an error bound that holds in probability, so Theorem 1.1 also allows us to develop conditions on the number n of samples that we need to control the size of the fluctuation. To obtain the sample complexity, assume that $n \leqslant p$, and select an error tolerance $\varepsilon \in (0,1)$. Then there is a constant $\mathrm{C}_{99\%}$ for which

$$n \geqslant \mathrm{C}_{99\%} \left[\frac{\|\boldsymbol{M}\|_{1\to 2}^2 \log p}{\varepsilon^2} + \frac{\|\boldsymbol{M}\| \log^2 p}{\varepsilon} \right] \frac{\|\boldsymbol{\Sigma}\|_{\max}}{\|\boldsymbol{\Sigma}\|} \quad \Longrightarrow \quad \left\| \boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma} \right\| \leqslant \varepsilon \|\boldsymbol{\Sigma}\| \tag{1.7}$$

with probability at least 99%. See the discussion after Theorem 3.1 for information about how to obtain sample complexity bounds that hold with higher probability.

1.3.6 *Example: The Banded Covariance Estimator.* Consider the banded covariance estimation problem in Section 1.2.1, with the mask

$$\boldsymbol{M} = \text{0–1 matrix with bandwidth } B.$$

The sample complexity bound (1.7) and the norm calculations from Section 1.3.1 demonstrate that

$$n \geqslant \mathrm{C} \left[\frac{B \log p}{\varepsilon^2} + \frac{B \log^2 p}{\varepsilon} \right] \cdot \frac{\|\boldsymbol{\Sigma}\|_{\max}}{\|\boldsymbol{\Sigma}\|} \tag{1.8}$$

is sufficient to obtain a relative spectral-norm error ε with constant probability. In particular, the condition $n \gtrsim B \log^2 p$ always ensures a constant relative error in (1.6). It follows that, when $B \ll p$, the variance of the estimator can be small, even when the number of samples is much smaller than the total number of variables.

1.3.7 *Comparison with Bounds of Levina and Vershynin.* The most natural point of comparison for Theorem 1.1 is the main theorem of Levina and Vershynin [24, Thm. 2.1]. Their result states that, for a centered normal random vector $\boldsymbol{x}$, the fluctuation in the masked sample covariance estimator satisfies

$$\mathbb{E} \left\| \boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma} \right\| \leqslant \mathrm{C} \left[\left(\frac{\|\boldsymbol{M}\|_{1\to 2}^2 \log^5 p}{n} \right)^{1/2} + \frac{\|\boldsymbol{M}\| \log^3 p}{n} \right] \|\boldsymbol{\Sigma}\| .$$

The associated sample complexity bound is

$$n \geqslant \mathrm{C} \left[\frac{\|\boldsymbol{M}\|_{1\to 2}^2 \log^5 p}{\varepsilon^2} + \frac{\|\boldsymbol{M}\| \log^3 p}{\varepsilon} \right] . \tag{1.9}$$

Our sample complexity bound (1.7) has a structure similar to (1.9), but several improvements are worth mentioning. First, the ratio of norms is a new feature in our estimate (1.7). The second improvement over (1.9), which has less conceptual significance, is the reduction of the number of logarithmic factors. Finally, our main result, Theorem 3.1 covers all centered distributions with four finite moments.

1.4 *Organization of the paper*

The rest of the paper is organized as follows. Section 2 introduces our notation and some preliminaries. Section 3 presents the main result for zero-mean distributions with finite fourth moments, together with its proof and the proof of Theorem 1.1. In Section 4, we deal with the technical estimates at the heart of the main result. Appendix A establishes the matrix concentration inequality we require.

2. Preliminaries

This section sets out the background material we need for the proof. Section 2.1 summarizes our notational conventions, and Section 2.2 describes some basic properties of the Schur product.

2.1 *Notation and Conventions*

In this paper, we work exclusively with real numbers. Plain italic letters always refer to scalars. Bold italic lowercase letters, such as $\boldsymbol{a}$, refer to column vectors. Bold italic uppercase letters, such as $\boldsymbol{A}$, denote matrices. All matrices in this work are square; the dimensions are determined by context. We write $\mathbf{0}$ for the zero matrix and $\mathbf{I}$ for the identity matrix. The matrix unit $\mathbf{E}_{ij}$ has a unit entry in the (i, j) position and zeros elsewhere.

The symbol * denotes the transpose operation on vectors and matrices. We use the term *self-adjoint* to refer to a matrix that satisfies $\boldsymbol{A} = \boldsymbol{A}^*$ to avoid confusion with symmetric random variables. Curly inequalities refer to the positive-semidefinite partial ordering on self-adjoint matrices: $\boldsymbol{A} \preccurlyeq \boldsymbol{B}$ if and only if $\boldsymbol{B} - \boldsymbol{A}$ is positive semidefinite.

The function $\operatorname{diag}(\cdot)$ maps a vector $\boldsymbol{a}$ to a matrix whose diagonal entries correspond with the entries of $\boldsymbol{a}$. When applied to a matrix, $\operatorname{diag}(\cdot)$ zeroes out the off-diagonal entries. We write $\operatorname{tr}(\cdot)$ for the trace. The symbol $\odot$ denotes the componentwise (i.e., Schur or Hadamard) product of two matrices.

We write $\|\cdot\|$ for both the ℓ_2 vector norm and the associated operator norm, which is usually called the *spectral norm*. The symbol $\|\cdot\|_q$ refers to the Schatten q-norm of a matrix:

$$\|\boldsymbol{A}\|_q := [\operatorname{tr}|\boldsymbol{A}|^q]^{1/q}$$

where $|\boldsymbol{A}| := (\boldsymbol{A}^*\boldsymbol{A})^{1/2}$. The norm $\|\cdot\|_\infty$ returns the maximum absolute entry of a vector, but we use a separate notation $\|\cdot\|_{\max}$ for the maximum absolute entry of a matrix. We also require the norm

$$\|\boldsymbol{A}\|_{1\to 2} := \max_j \left(\sum\nolimits_i |a_{ij}|^2\right)^{1/2}.$$

The notation reflects the fact that this is the natural norm for linear maps from ℓ_1 into ℓ_2.

We reserve the symbol ξ for a *Rademacher random variable*, which takes the two values ± 1 with equal probability. We also assume that all random variables are sufficiently regular that we are justified in computing expectations, interchanging limits, and so forth.

2.2 *Facts about the Schur Product*

The proof depends on some basic properties of Schur products. The first result is a simple but useful algebraic identity. For each square matrix $\boldsymbol{A}$ and each conforming vector $\boldsymbol{x}$,

$$\boldsymbol{A} \odot \boldsymbol{x}\boldsymbol{x}^* = \operatorname{diag}(\boldsymbol{x})\,\boldsymbol{A}\,\operatorname{diag}(\boldsymbol{x}). \tag{2.1}$$

The second result states that the Schur product with a positive-semidefinite matrix is order preserving. That is, for a fixed positive-semidefinite matrix $\boldsymbol{A}$,

$$\boldsymbol{B}_1 \preccurlyeq \boldsymbol{B}_2 \quad \text{implies} \quad \boldsymbol{A} \odot \boldsymbol{B}_1 \preccurlyeq \boldsymbol{A} \odot \boldsymbol{B}_2. \tag{2.2}$$

This property follows from Schur's theorem [16, Thm. 7.5.3], which states that the Schur product of two positive-semidefinite matrices remains positive semidefinite.

3. Masked Covariance Estimation

In this section, we state and prove detailed error estimates for masked covariance estimation of a general distribution with finite fourth moments. Section 3.1 defines two concentration parameters that measure the spread of the distribution. We present the main theorem and a short discussion in Sections 3.2 and 3.3. In Section 3.4, we show how to derive Theorem 1.1, the result for Gaussian distributions. The proof of the main result appears in Section 3.5.

3.1 *Concentration Parameters*

The effectiveness of the masked sample covariance estimator depends on the concentration properties of the distribution of $\boldsymbol{x}$. Let us introduce two quantities that measure different facets of the variation of the random vector.

For $r \geqslant 1$, the *rth diagonal moment* $\mu_r(\boldsymbol{x})$ of the distribution is defined to be the maximum L_r norm of a single component of the vector:

$$\mu_r(\boldsymbol{x}) := \max_i \left(\mathbb{E}\,|X_i|^r\right)^{1/r}. \tag{3.1}$$

In other words, μ_r gives us uniform control on the rth moment of each component of $\boldsymbol{x}$.

We also require some information about the spread of the distribution in all directions. Define the *uniform fourth moment* $\nu(\boldsymbol{x})$ by the formula

$$\nu(\boldsymbol{x}) := \sup_{\|\boldsymbol{u}\|=1} \left(\mathbb{E}\,|\boldsymbol{u}^*\boldsymbol{x}|^4\right)^{1/4}. \tag{3.2}$$

The uniform fourth moment measures how much the worst marginal varies.

Note that both $\mu_r(\boldsymbol{x})$ and $\nu(\boldsymbol{x})$ have the same homogeneity as the random vector $\boldsymbol{x}$. (This property is sometimes expressed by saying that the quantities have the same dimension, the same units, or the same scaling.) As a consequence, the quantities $\mu_r(\boldsymbol{x})\nu(\boldsymbol{x})$ and $\mu_r^2(\boldsymbol{x})$ have the same homogeneity as the covariance matrix $\boldsymbol{\Sigma}$. In the sequel, we abbreviate $\mu_r := \mu_r(\boldsymbol{x})$ and $\nu := \nu(\boldsymbol{x})$ whenever the distribution of the random vector $\boldsymbol{x}$ is clear.

3.2 *Main Result for Masked Covariance Estimation*

The following theorem provides detailed information about the variance of the error in the masked sample covariance estimator for a zero-mean distribution with finite fourth moments.

THEOREM 3.1 (The Masked Sample Covariance Estimator) Fix a $p \times p$ symmetric mask matrix $\boldsymbol{M}$, where $p \geqslant 3$. Suppose that $\boldsymbol{x}$ is a random vector in $\mathbb{R}^p$ with mean zero. Define the covariance matrix $\boldsymbol{\Sigma}$ and the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_n$ as in (1.1) and (1.2). Then the variance in the masked sample covariance estimator satisfies

$$\left[\mathbb{E}\left\|\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M}\odot\boldsymbol{\Sigma}\right\|^2\right]^{1/2} \leqslant \sqrt{\frac{8\mathrm{e}\log p}{n}}\cdot\|\boldsymbol{M}\|_{1\to 2}\cdot\mu_4\nu + \frac{8\mathrm{e}\log p}{n}\cdot\|\boldsymbol{M}\|\cdot\left[\mathbb{E}\max_i\|\boldsymbol{x}_i\|_\infty^4\right]^{1/2}. \tag{3.3}$$

Furthermore, the expected maximum satisfies the bound

$$\left[\mathbb{E}\max_i\|\boldsymbol{x}_i\|_\infty^4\right]^{1/2} \leqslant \inf_{r\geqslant 1}(np)^{1/2r}\cdot\mu_{4r}^2. \tag{3.4}$$

The diagonal moment μ_r and the uniform fourth moment ν are defined in (3.1) and (3.2).

In Section 3.3, we offer a short discussion of this result. Afterward, in Section 3.4, we specialize the result to Gaussian distributions, which establishes Theorem 1.1 of the Introduction. The proof of Theorem 3.1 appears below in Section 3.5.

3.3 *Discussion*

Theorem 3.1 has a wider scope that most of the results in the literature on sparse covariance estimation. Indeed, we allow completely general masks, and the bound is valid for any distribution with finite fourth moments. When we specialize the result to the Gaussian case, we obtain an improvement over prior work [24, Thm. 2.1]. Even so, our argument, which is based on a matrix moment inequality, is very direct.

For simplicity, we have presented Theorem 3.1 as a bound on the variance of the masked sample covariance estimator. A refinement of the same argument allows us to compute higher moments of the error, which in turn yield polynomial tail bounds via Markov's inequality. When the distribution of the random vector $\boldsymbol{x}$ is subgaussian, this method even yields exponential tail bounds.

REMARK 3.1 (Alternative Arguments) The proof of Theorem 3.1 is based on a new matrix moment inequality. We can obtain similar results using other matrix concentration inequalities that appear in the literature. In particular, the matrix Rosenthal inequality [27, Cor. 7.5] leads to a very similar bound. The initial version of this manuscript [9] uses the matrix Bernstein inequality [38] to develop a version of Theorem 3.1 for a subgaussian random vector $\boldsymbol{x}$.

3.4 *Specialization to Gaussian Distributions*

It is natural to apply Theorem 3.1 to study the performance of masked covariance estimation for a zero-mean Gaussian random vector. In this case, the covariance matrix determines the distribution completely, so we can obtain a more transparent statement that does not involve the concentration parameters. Theorem 1.1 follows from these considerations.

Proof of Theorem 1.1 from Theorem 3.1. First, we compute the $(2r)$th diagonal moment $\mu_{2r}(\boldsymbol{x})$ for $r \geqslant 1$. Observe that the ith component X_i of the vector $\boldsymbol{x}$ is a centered normal random variable with variance σ_{ii}, where σ_{ii} denotes the ith diagonal entry of $\boldsymbol{\Sigma}$. Using the standard expression for the $(2r)$th moment of a normal random variable, we obtain

$$\mathbb{E}\,|X_i|^{2r} = \frac{(2r)!}{2^r\, r!} \cdot \sigma_{ii}^r \leqslant r^r \cdot \sigma_{ii}^r.$$

This bound is valid for each real number $r \geqslant 1$. Therefore, taking the rth root, we reach

$$\mu_{2r}^2 \leqslant r \cdot \max_i \sigma_{ii} = r\,\|\boldsymbol{\Sigma}\|_{\max}. \tag{3.5}$$

The identity holds because the maximum entry of a positive-definite matrix occurs on its diagonal. In particular, we see that

$$\mu_4 \leqslant \sqrt{2}\,\|\boldsymbol{\Sigma}\|_{\max}^{1/2}. \tag{3.6}$$

Next, we instantiate the bound (3.4) on the expected maximum. Choose $4r = 2\log(np)$ to reach

$$\left[\mathbb{E}\max_i \|\boldsymbol{x}_i\|_\infty^4\right]^{1/2} \leqslant \mathrm{e}\,\mu_{2\log(np)}^2 \leqslant \mathrm{e}\,\log(np) \cdot \|\boldsymbol{\Sigma}\|_{\max}, \tag{3.7}$$

owing to (3.5).

Finally, we bound the uniform fourth moment $\nu(\boldsymbol{x})$. Fix a unit vector $\boldsymbol{u}$. The distribution of the marginal $\boldsymbol{u}^*\boldsymbol{x}$ is Gaussian with mean zero. To compute the variance $\sigma_{\boldsymbol{u}}^2$ of the marginal, we write $\boldsymbol{x} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{g}$, where $\boldsymbol{g}$ is a standard Gaussian vector. Then

$$\sigma_{\boldsymbol{u}}^2 = \mathbb{E}\,|\boldsymbol{u}^*\boldsymbol{x}|^2 = \mathbb{E}\,|\boldsymbol{u}^*(\boldsymbol{\Sigma}^{1/2}\boldsymbol{g})|^2 = \boldsymbol{u}^*\boldsymbol{\Sigma}^{1/2}(\mathbb{E}\,\boldsymbol{g}\boldsymbol{g}^*)\boldsymbol{\Sigma}^{1/2}\boldsymbol{u} = \boldsymbol{u}^*\boldsymbol{\Sigma}\boldsymbol{u} \leqslant \|\boldsymbol{\Sigma}\|\,.$$

The fourth moment of a Gaussian variable equals three times its squared variance, so

$$\mathbb{E}\,|\boldsymbol{u}^*\boldsymbol{x}|^4 = 3\sigma_{\boldsymbol{u}}^4 \leqslant 3\,\|\boldsymbol{\Sigma}\|^2\,.$$

We conclude that the uniform fourth moment satisfies

$$\nu(\boldsymbol{x}) = \sup_{\|\boldsymbol{u}\|=1} (\mathbb{E}\,|\boldsymbol{u}^*\boldsymbol{x}|^4)^{1/4} \leqslant 3^{1/4}\,\|\boldsymbol{\Sigma}\|^{1/2}. \tag{3.8}$$

To complete the proof of Theorem 1.1, substitute the bounds (3.6), (3.7), and (3.8) into the inequality (3.3) from Theorem 3.1. □

3.5 *Proof of Theorem 3.1*

The proof of Theorem 3.1 proceeds in several short steps. First, we write the variance of the estimator as a sum of independent random matrices, and we use symmetrization to simplify the expression. Second, we apply a matrix moment inequality to bound the variance of the estimator in terms of the spectral norm of a matrix variance and the maximum spectral norm of the summands. Finally, some short computations, which appear in Section 4, yield bounds for the remaining terms.

3.5.1 *Symmetrization.*

We begin by assigning a name to the quantity of interest:

$$E := \mathbb{E}\,\big\|\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}\big\|.$$

The random matrix inside the norm has a natural expression as a sum of independent, centered random matrices. To see why, substitute the definitions (1.1) and (1.2) of the population covariance matrix $\boldsymbol{\Sigma}$ and the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_n$ to obtain

$$E = \frac{1}{n}\cdot\mathbb{E}\,\Big\|\sum\nolimits_{i=1}^n (\boldsymbol{M} \odot \boldsymbol{x}_i\boldsymbol{x}_i^* - \mathbb{E}\,\boldsymbol{M} \odot \boldsymbol{x}_i\boldsymbol{x}_i^*)\Big\|\,.$$

The standard symmetrization method [23, Lem. 6.3] yields the bound

$$E \leqslant \frac{2}{n}\cdot\mathbb{E}\,\Big\|\sum\nolimits_{i=1}^n \xi_i(\boldsymbol{M} \odot \boldsymbol{x}_i\boldsymbol{x}_i^*)\Big\|\,. \tag{3.9}$$

Here, $\{\xi_i\}$ is a sequence of independent Rademacher random variables that is also independent from the sequence $\{\boldsymbol{x}_i\}$ of samples. The advantage of the expression (3.9) is that each Schur product involves a rank-one matrix, which greatly simplifies our computations.

3.5.2 *The Spectral Norm of an Independent Sum.* The main technical tool in this paper is a bound for the second moment of the spectral norm of a sum of independent, symmetric random matrices.

THEOREM 3.2 (Matrix Second Moment Inequality) Assume that $p \geqslant 3$. Consider a finite sequence $\{\boldsymbol{Y}_i\}$ of independent, symmetric, random, self-adjoint matrices with dimension $p \times p$. Then

$$\left[\mathbb{E}\left\|\sum\nolimits_i \boldsymbol{Y}_i\right\|^2\right]^{1/2} \leqslant \sqrt{2e\log p}\cdot\left\|\left[\sum\nolimits_i \mathbb{E}\,\boldsymbol{Y}_i^2\right]^{1/2}\right\| + 4e\log p\cdot\left[\mathbb{E}\max\nolimits_i \|\boldsymbol{Y}_i\|^2\right]^{1/2}.$$

Theorem 3.2 reduces the challenging problem of bounding the second moment of the spectral norm to two simpler calculations. We interpret the first term as the variance of a sum of independent, symmetric random matrices. The second term measures the typical size of the largest summand. The result is new in the form that we present it, but it has strong precedents in the literature. See Appendix A for the proof and a discussion of related work.

With Theorem 3.2 at hand, it is straightforward to bound (3.9). We reach

$$\begin{aligned} E &\leqslant \frac{1}{n}\sqrt{8e\log p}\left\|\left[\sum\nolimits_i \mathbb{E}(\boldsymbol{M}\odot\boldsymbol{x}_i\boldsymbol{x}_i^*)^2\right]^{1/2}\right\| + \frac{8e\log p}{n}\left[\mathbb{E}\max\nolimits_i \|\boldsymbol{M}\odot\boldsymbol{x}_i\boldsymbol{x}_i^*\|^2\right]^{1/2} \\ &= \sqrt{\frac{8e\log p}{n}}\left\|\mathbb{E}(\boldsymbol{M}\odot\boldsymbol{x}\boldsymbol{x}^*)^2\right\|^{1/2} + \frac{8e\log p}{n}\left[\mathbb{E}\max\nolimits_i \|\boldsymbol{M}\odot\boldsymbol{x}_i\boldsymbol{x}_i^*\|^2\right]^{1/2}. \end{aligned} \tag{3.10}$$

The second line follows from the identical distribution of the summands.

3.5.3 *The Matrix Variance and the Maximum Spectral Norm.* All that remains is to calculate the matrix variance that appears in the first term of (3.10) and the expected maximum norm that appears in the second term. Lemma 4.1 demonstrates that

$$\mathbb{E}(\boldsymbol{M}\odot\boldsymbol{x}\boldsymbol{x}^*)^2 \preccurlyeq \mu_4^2\nu^2\cdot\|\boldsymbol{M}\|_{1\to 2}^2\cdot\mathbf{I}. \tag{3.11}$$

The concentration parameters μ_r and ν that characterize $\boldsymbol{x}$ are defined in (3.1) and (3.2). Lemma 4.2 provides a simple deterministic estimate for the remaining Schur product:

$$\|\boldsymbol{M}\odot\boldsymbol{x}\boldsymbol{x}^*\| \leqslant \|\boldsymbol{M}\|\,\|\boldsymbol{x}\|_\infty^2. \tag{3.12}$$

Introduce the matrix variance bound (3.11) and the Schur product bound (3.12) into (3.10) to obtain

$$E \leqslant \sqrt{\frac{8e\log p}{n}}\cdot\|\boldsymbol{M}\|_{1\to 2}\cdot\mu_4\nu + \frac{8e\log p}{n}\cdot\|\boldsymbol{M}\|\cdot\left[\mathbb{E}\max\nolimits_i \|\boldsymbol{x}_i\|_\infty^4\right]^{1/2}. \tag{3.13}$$

To incorporate the semidefinite second moment bound, we have used the fact that the spectral norm is monotone with respect to the order $\preccurlyeq$ on the set of positive semidefinite matrices. This is the first claim in Theorem 3.1.

To establish the remaining claim (3.4), we need a bound for the expected maximum of $\|\boldsymbol{x}_i\|_\infty^4$. Lemma 4.3 states that

$$\mathbb{E}\max\nolimits_i \|\boldsymbol{x}_i\|_\infty^4 \leqslant \inf_{r\geqslant 1}\,(np)^{1/r}\cdot\mu_{4r}^4. \tag{3.14}$$

This observation completes the proof.

4. Computing the Matrix Variance and the Maximum Spectral Norm

In this section, we complete the calculations that stand at the center of Theorem 3.1.

4.1 *A Semidefinite Bound for the Matrix Variance*

First, we study the matrix variance $\mathbb{E}(\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2$. This calculation requires some insight, and it is the main novelty in our proof. The key idea is that the monotonicity (2.2) of the Schur product allows us to replace one factor in the product by a scalar matrix. This act of diagonalization simplifies the estimate tremendously because we erase the off-diagonal entries when we take the Schur product with an identity matrix.

LEMMA 4.1 (Matrix Variance Bound) Fix a self-adjoint $p \times p$ matrix $\boldsymbol{M}$. Let $\boldsymbol{x} = (X_1, \dots, X_p)^*$ be a random vector. Then

$$\mathbb{E}(\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2 \preccurlyeq \mu_4^2 \nu^2 \, \|\boldsymbol{M}\|_{1\to 2}^2 \cdot \mathbf{I}.$$

The concentration parameters μ_4 and ν are defined in (3.1) and (3.2).

Proof. To begin, we perform some algebraic manipulations to consolidate the randomness. The Schur product identity (2.1) implies that

$$\begin{aligned}(\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2 &= (\operatorname{diag}(\boldsymbol{x})\boldsymbol{M}\operatorname{diag}(\boldsymbol{x}))^2 \\ &= \operatorname{diag}(\boldsymbol{x})(\boldsymbol{M}\operatorname{diag}(\boldsymbol{x})^2\boldsymbol{M})\operatorname{diag}(\boldsymbol{x}) = (\boldsymbol{M}\operatorname{diag}(\boldsymbol{x})^2\boldsymbol{M}) \odot \boldsymbol{x}\boldsymbol{x}^*.\end{aligned}$$

Rewrite the diagonal matrix as a linear combination of matrix units: $\operatorname{diag}(\boldsymbol{x})^2 = \sum_i X_i^2 \mathbf{E}_{ii}$. The bilinearity of the Schur product now yields

$$(\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2 = \left[\boldsymbol{M}\left(\textstyle\sum_i X_i^2 \mathbf{E}_{ii}\right)\boldsymbol{M}\right] \odot \boldsymbol{x}\boldsymbol{x}^* = \textstyle\sum_i (\boldsymbol{M}\mathbf{E}_{ii}\boldsymbol{M}) \odot (X_i^2\, \boldsymbol{x}\boldsymbol{x}^*).$$

Take the expectation of this expression to reach

$$\mathbb{E}(\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2 = \textstyle\sum_i (\boldsymbol{M}\mathbf{E}_{ii}\boldsymbol{M}) \odot [\mathbb{E}(X_i^2\, \boldsymbol{x}\boldsymbol{x}^*)]. \tag{4.1}$$

Next, we invoke the monotonicity (2.2) of the Schur product to make a diagonal estimate for each summand in (4.1):

$$(\boldsymbol{M}\mathbf{E}_{ii}\boldsymbol{M}) \odot [\mathbb{E}(X_i^2\boldsymbol{x}\boldsymbol{x}^*)] \preccurlyeq \lambda_{\max}(\mathbb{E}(X_i^2\boldsymbol{x}\boldsymbol{x}^*)) \cdot (\boldsymbol{M}\mathbf{E}_{ii}\boldsymbol{M}) \odot \mathbf{I}.$$

The Rayleigh–Ritz variational formula [3, Cor. III.1.2] allows us to write the maximum eigenvalue as a supremum. Thus,

$$\begin{aligned}\lambda_{\max}(\mathbb{E}(X_i^2\boldsymbol{x}\boldsymbol{x}^*)) &= \sup_{\|\boldsymbol{u}\|=1} \boldsymbol{u}^*\,[\mathbb{E}(X_i^2\,\boldsymbol{x}\boldsymbol{x}^*)]\,\boldsymbol{u} = \sup_{\|\boldsymbol{u}\|=1} \mathbb{E}\left[X_i^2\,|\boldsymbol{u}^*\boldsymbol{x}|^2\right] \\ &\leqslant \sup_{\|\boldsymbol{u}\|=1} (\mathbb{E}X_i^4)^{1/2}\,(\mathbb{E}\,|\boldsymbol{u}^*\boldsymbol{x}|^4)^{1/2} \leqslant \mu_4^2\nu^2.\end{aligned}$$

The first inequality is Cauchy–Schwarz. The final inequality follows from the definitions (3.1) and (3.2) of the concentration parameters. Combine the last two displays to obtain

$$(\boldsymbol{M}\mathbf{E}_{ii}\boldsymbol{M}) \odot [\mathbb{E}(X_i^2\boldsymbol{x}\boldsymbol{x}^*)] \preccurlyeq \mu_4^2\nu^2 \cdot (\boldsymbol{M}\mathbf{E}_{ii}\boldsymbol{M}) \odot \mathbf{I}. \tag{4.2}$$

To complete our bound for the variance, we introduce (4.2) into (4.1), which delivers

$$\mathbb{E}(\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2 \preccurlyeq \mu_4^2\nu^2 \cdot \boldsymbol{M}^2 \odot \mathbf{I} = \mu_4^2\nu^2 \cdot \operatorname{diag}(\boldsymbol{M}^2)$$

We can control a positive-semidefinite diagonal matrix using only its maximum entry:

$$\mathbb{E}(\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2 \preccurlyeq \mu_4^2\nu^2 \cdot \max_i(\boldsymbol{M}^2)_{ii} \cdot \mathbf{I} = \mu_4^2\nu^2\,\|\boldsymbol{M}\|_{1\to 2}^2 \cdot \mathbf{I}$$

The second relation follows from the fact that the diagonal entries of $\boldsymbol{M}^2$ list the squared ℓ_2 norms of the columns of $\boldsymbol{M}$, while $\|\boldsymbol{M}\|_{1\to 2}$ computes the maximum ℓ_2 norm of a column of $\boldsymbol{M}$. □

4.2 *Norm Bound for a Schur Product*

Next, we present a simple norm bound for the Schur product $\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*$.

LEMMA 4.2 (Norm Bound for a Schur Product) Let $\boldsymbol{M}$ be a $p \times p$ self-adjoint matrix, and let $\boldsymbol{x}$ be a vector in $\mathbb{R}^p$. Then

$$\|\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*\| \leqslant \|M\| \, \|\boldsymbol{x}\|_\infty^2 .$$

Proof. The Hadamard product identity (2.1) yields

$$\|\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*\| = \|\operatorname{diag}(\boldsymbol{x}) \boldsymbol{M} \operatorname{diag}(\boldsymbol{x})\| \leqslant \|\operatorname{diag}(\boldsymbol{x})\| \, \|\boldsymbol{M}\| \, \|\operatorname{diag}(\boldsymbol{x})\| = \|\boldsymbol{M}\| \, \|\boldsymbol{x}\|_\infty^2 .$$

The inequality follows from the submultiplicativity of the spectral norm. □

4.3 *The Expected Maximum Entry among the Sample Vectors*

Finally, we develop a basic estimate on the expected maximum entry that appears in any of the sample vectors.

LEMMA 4.3 (The Expected Maximum) Consider an i.i.d. sequence $\{\boldsymbol{x}_i\}_{i=1}^n$ of random vectors in $\mathbb{R}^p$. Then

$$\left[\mathbb{E} \max_i \|\boldsymbol{x}_i\|_\infty^4\right]^{1/4} \leqslant \inf_{r \geqslant 1} (np)^{1/4r} \cdot \mu_{4r}.$$

The diagonal moment parameter μ_{4r} is defined in (3.1).

Proof. For any $r \geqslant 1$, Jensen's inequality yields

$$\left[\mathbb{E} \max_i \|\boldsymbol{x}_i\|_\infty^4\right]^{1/4} \leqslant \left[\mathbb{E} \max_i \|\boldsymbol{x}_i\|_\infty^{4r}\right]^{1/4r} \leqslant \left[\sum_{i=1}^n \sum_{j=1}^p \mathbb{E} \left|X_{ij}\right|^{4r}\right]^{1/4r} \leqslant (np)^{1/4r} \mu_{4r}.$$

We have written X_{ij} for the jth entry of $\boldsymbol{x}_i$ and invoked the definition of the diagonal moment μ_{4r}. □

A. The Matrix Moment Inequality

This appendix contains a proof of Theorem 3.2, the matrix moment inequality that animates our argument. It costs us no additional energy to prove a result that holds for all moments.

THEOREM A.1 (Matrix Moment Inequality) Assume that $p \geqslant 3$.

1. Suppose that $q \geqslant 1$, and fix $r \geqslant \max\{q, 2\log p\}$. Consider a finite sequence $\{\boldsymbol{W}_i\}$ of independent, random, positive-semidefinite matrices with dimension $p \times p$. Then

$$\left[\mathbb{E} \left\|\textstyle\sum_i \boldsymbol{W}_i\right\|^q\right]^{1/q} \leqslant \left[\left\|\textstyle\sum_i \mathbb{E}\, \boldsymbol{W}_i\right\|^{1/2} + 2\sqrt{\mathrm{e}\, r} \left(\mathbb{E} \max_i \|\boldsymbol{W}_i\|^q\right)^{1/2q}\right]^2 \tag{A.1}$$

2. Suppose that $q \geqslant 2$, and fix $r \geqslant \max\{q, 2\log p\}$. Consider a finite sequence $\{\boldsymbol{Y}_i\}$ of independent, symmetric, random, self-adjoint matrices with dimension $p \times p$. Then

$$\left[\mathbb{E} \left\|\textstyle\sum_i \boldsymbol{Y}_i\right\|^q\right]^{1/q} \leqslant \sqrt{\mathrm{e}\, r} \left\|\left(\textstyle\sum_i \mathbb{E}\, \boldsymbol{Y}_i^2\right)^{1/2}\right\| + 2\mathrm{e}\, r \left(\mathbb{E} \max_i \|\boldsymbol{Y}_i\|^q\right)^{1/q}. \tag{A.2}$$

Theorem 3.2 follows from (A.2) when we select $q = 2$ and $r = 2\log p$. We establish Theorem A.1 below after we provide some comments, preliminary results, and historical background.

Theorem A.1 shows that the moments of a spectral norm of a sum are controlled by two different quantities. The first term in (A.1) is a matrix mean, while the first term in (A.2) is a matrix variance. These terms reflect the size of moderate deviations, and they depend only weakly on the order q of the moment. The second term measures the size of the largest summand, and it controls the large deviation behavior of the sum. These bounds are related to the matrix Rosenthal inequality [22, 27], and they can be viewed as the moment inequality underlying the matrix Bernstein inequality [38, Thm. 1.4].

A.1 *Matrix Khintchine Inequality*

The main ingredient in the proof of Theorem A.1 is the matrix Khintchine inequality. To our knowledge, this result is the earliest matrix moment inequality.

PROPOSITION A.2 (Matrix Khintchine Inequality) Suppose that $r \geqslant 2$. Consider a finite sequence $\{\boldsymbol{A}_i\}$ of deterministic, self-adjoint matrices. Then

$$\left[\mathbb{E}\left\|\sum\nolimits_i \xi_i \boldsymbol{A}_i\right\|_r^r\right]^{1/r} \leqslant \sqrt{r}\left\|\left[\sum\nolimits_i \boldsymbol{A}_i^2\right]^{1/2}\right\|_r.$$

The sequence $\{\xi_i\}$ consists of independent Rademacher random variables.

Lust-Picquard established the first version of the matrix Khintchine inequality in [25], with a weaker estimate for the constant. Her subsequent paper with Pisier [26] contains important extensions and refinements. Buchholz obtained sharp constants for even r in [7]. The recent work [27, Sec. 7] contains what may be the easiest proof of Proposition A.2; it yields the near-optimal bound $\sqrt{r}$ for the constant.

A.2 *Historical Background on Matrix Concentration Inequalities*

Research on matrix moment inequalities contains several strands that date back to the late 1990s. The paper [31] of Pisier and Xu initiated the field of noncommutative martingale inequalities. This literature contains many powerful moment bounds that can also be used to study sums of independent random matrices [19–22]. An early application of this theory appeared in Rudelson's paper [34], which uses the matrix Khintchine inequality to obtain a sample complexity bound for classical covariance estimation. Many authors in computer science, mathematical signal processing, and other areas adapted Rudelson's method [35, 36, 39].

There is a parallel line of work that develops exponential moment inequalities for sums of random matrices. This research was initiated in the paper Ahlswede–Winter [1] and continued in a variety of other works [14, 27, 30, 32, 37, 38]. Over the last few years, these results have started to see wide application.

As it is stated, Theorem A.1 seems to be new. Nevertheless, the result is substantially similar to some previous matrix concentration inequalities that appear in the literature. Indeed, we have adapted the argument from Rudelson's paper [34], the refinements of Rudelson's work in [35, 36], and the recent proofs of two matrix Rosenthal inequalities [22, 27].

A.3 *Proof of the Matrix Moment Inequality, Part I*

We prove Theorem A.1 in two steps. First, we establish the inequality (A.1) for positive matrices. In the second stage, we extend this result to obtain the bound (A.2) for general matrices.

To begin, we introduce the quantity of interest:

$$E_q^2 := \left[\mathbb{E}\left\|\sum_i \boldsymbol{W}_i\right\|^q\right]^{1/q} \leqslant 2\left[\mathbb{E}\left\|\sum_i \xi_i \boldsymbol{W}_i\right\|^q\right]^{1/q} + \left\|\sum_i \mathbb{E}\,\boldsymbol{W}_i\right\|. \tag{A.3}$$

The inequality follows when we center the sum and apply the standard symmetrization result [23, Lem. 6.3]. The sequence $\{\xi_i\}$ consists of independent Rademacher random variables that are also independent from the sequence $\{\boldsymbol{P}_i\}$.

Let us focus on the first term on the right-hand side of (A.3):

$$F_q := \left[\mathbb{E}\left\|\sum_i \xi_i \boldsymbol{W}_i\right\|^q\right]^{1/q} \leqslant \left[\mathbb{E}_{\boldsymbol{W}_i}\left(\mathbb{E}_{\xi_i}\left\|\sum_i \xi_i \boldsymbol{W}_i\right\|_r^r\right)^{q/r}\right]^{1/q}.$$

The inequality holds because the Schatten r-norm dominates the spectral norm, and we have used the fact $q \leqslant r$ to apply Jensen's inequality to the inner expectation. An application of the matrix Khintchine inequality, Proposition A.2, delivers the bound

$$F_q \leqslant \sqrt{r}\left[\mathbb{E}\left\|\left(\sum_i \boldsymbol{W}_i^2\right)^{1/2}\right\|_r^q\right]^{1/q}.$$

We proceed through a short chain of inequalities to complete the estimate:

$$\begin{aligned} F_q &\leqslant \sqrt{\mathrm{e}r}\left[\mathbb{E}\left\|\sum_i \boldsymbol{W}_i^2\right\|^{q/2}\right]^{1/q} \\ &\leqslant \sqrt{\mathrm{e}r}\left[\mathbb{E}\left(\max_i \|\boldsymbol{W}_i\|^{q/2} \cdot \left\|\sum_i \boldsymbol{W}_i\right\|^{q/2}\right)\right]^{1/q} \\ &\leqslant \sqrt{\mathrm{e}r}\left[\mathbb{E}\max_i \|\boldsymbol{W}_i\|^q\right]^{1/2q} \cdot \left[\mathbb{E}\left\|\sum_i \boldsymbol{W}_i\right\|^q\right]^{1/2q}. \end{aligned}$$

In the preceding calculation, we first replace the Schatten r-norm by the spectral norm, which results in a loss of at most $p^{1/r} \leqslant \sqrt{\mathrm{e}}$. Since $\boldsymbol{W}_i$ is positive semidefinite, we can make the bound $\boldsymbol{W}_i^2 \preccurlyeq \|\boldsymbol{W}_i\| \cdot \boldsymbol{W}_i$ and invoke the monotonicity of the spectral norm on the positive-semidefinite cone to draw off the maximum norm achieved by any one of the summands. The last inequality is Cauchy–Schwarz. We conclude that

$$F_q \leqslant \sqrt{\mathrm{e}r}\left[\mathbb{E}\max_i \|\boldsymbol{W}_i\|^q\right]^{1/2q} \cdot E_q \tag{A.4}$$

by identifying a copy of E_q.

To complete the argument, introduce (A.4) into the inequality (A.3):

$$E_q^2 \leqslant 2\sqrt{\mathrm{e}r}\left[\mathbb{E}\max_i \|\boldsymbol{W}_i\|^q\right]^{1/2q} \cdot E_q + \left\|\sum_i \mathbb{E}\,\boldsymbol{W}_i\right\|.$$

Solutions to the quadratic inequality $x^2 \leqslant ax + b$ satisfy $x \leqslant a + \sqrt{b}$. Therefore,

$$E_q \leqslant 2\sqrt{\mathrm{e}r}\left[\mathbb{E}\max_i \|\boldsymbol{W}_i\|^q\right]^{1/2q} + \left\|\sum_i \mathbb{E}\,\boldsymbol{W}_i\right\|^{1/2}.$$

This estimate coincides with the bound (A.1).

A.4 *Proof of the Matrix Moment Inequality, Part II*

To establish the second bound (A.2), we apply the matrix Khintchine inequality to obtain a bound involving positive-semidefinite random matrices, and then we invoke our first result (A.1). Indeed, since the summands $\boldsymbol{Y}_i$ are symmetric random variables,

$$\left[\mathbb{E}\left\|\sum\nolimits_i \boldsymbol{Y}_i\right\|^q\right]^{1/q} = \left[\mathbb{E}\left\|\sum\nolimits_i \xi_i \boldsymbol{Y}_i\right\|^q\right]^{1/q} \leqslant \left[\mathbb{E}_{\boldsymbol{Y}_i}\left(\mathbb{E}_{\varepsilon_i}\left\|\sum\nolimits_i \xi_i \boldsymbol{Y}_i\right\|_r^r\right)^{q/r}\right]^{1/q}.$$

The inequality follows from the same considerations as in the proof of (A.2). Invoke Proposition A.2 to obtain

$$\left[\mathbb{E}\left\|\sum\nolimits_i \boldsymbol{Y}_i\right\|^q\right]^{1/q} \leqslant \sqrt{r}\left[\mathbb{E}\left\|\left(\sum\nolimits_i \boldsymbol{Y}_i^2\right)^{1/2}\right\|_r^q\right]^{1/q} \leqslant \sqrt{\mathrm{e}\,r}\left[\mathbb{E}\left\|\sum\nolimits_i \boldsymbol{Y}_i^2\right\|^{q/2}\right]^{1/q}.$$

In the second step, we have replaced the Schatten r-norm with the spectral norm; the third step follows from Jensen's inequality. The resulting expression involves a sum of independent, random positive-semidefinite matrices. Since $q/2 \geqslant 1$, we can apply (A.1) with $\boldsymbol{W}_i = \boldsymbol{Y}_i^2$ to reach the conclusion (A.2).

References

[1] AHLSWEDE, R. & WINTER, A. (2002) Strong converse for identification via quantum channels.. *IEEE Trans. Inform. Theory*, **48**(3), 569–579.

[2] BAI, Z. D. & SILVERSTEIN, J. W. (2010) *Spectral Analysis of Large-Dimensional Random Matrices*. Springer, New York, NY.

[3] BHATIA, R. (1997) *Matrix Analysis*. Springer, New York, NY.

[4] BICKEL, P. J. & LEVINA, E. (2008a) Covariance regularization by thresholding. *Ann. Statist.*, **36**(6), 2577–2604.

[5] ——— (2008b) Regularized estimation of large covariance matrices. *Ann. Statist.*, **36**(1), 199–227.

[6] BUCHHOLZ, A. (2001) Operator Khintchine inequality in non-commutative probability. *Math. Ann.*, **319**, 1–16.

[7] ——— (2005) Optimal constants in Khintchine-type inequalities for Fermions, Rademachers and q-Gaussian operators. *Bull. Pol. Acad. Sci. Math.*, **53**(3), 315–321.

[8] CAI, T. T., ZHANG, C.-H. & ZHOU, H. H. (2010) Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.,*, **38**(4), 2118–2144.

[9] CHEN, R. Y., GITTENS, A. & TROPP, J. A. (2012) The masked sample covariance estimator: An analysis via the matrix Laplace transform. ACM Report 2012-01, California Inst. Tech., Pasadena, CA.

[10] DE LA PEÑA, V. H. & GINÉ, E. (1999) *Decoupling: From Dependence to Independence*. Springer, Berlin.

[11] EL KAROUI, N. (2008) Operator norm consistent estimation of large dimensional sparse covariance matrices. *Ann. Statist.*, **36**(6), 2717–2756.

[12] FREEDMAN, D. A. (2005) *Statistical Models: Theory and Practice*. Cambridge Univ. Press, Cambridge.

[13] FURRER, R. & BENGTSSON, T. (2007) Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *J. Multivar. Anal.*, **98**(2), 227–255.

[14] GROSS, D. (2011) Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory*, **57**(3), 1548–1566.

[15] HORN, R. A. & JOHNSON, C. R. (1985) *Matrix Analysis*. Cambridge Univ. Press, Cambridge.

[16] ——— (1994) *Topics in Matrix Analysis*. Cambridge Univ. Press, Cambridge.

[17] JOHNSON, R. A. & WICHERN, D. W. (2002) *Applied Multivariate Statistical Analysis*. Prentice Hall, Englewood Cliffs, NJ, 6th edn.

[18] JOLLIFFE, I. T. (2002) *Principal Component Analysis*. Springer, New York, NY.

[19] JUNGE, M. & XU, Q. (2003) Noncommutative Burkholder/Rosenthal Inequalities. *Ann. Probab.*, **31**(2), 948–995.

[20] ——— (2005) On the best constants in some non-commutative martingale inequalities. *Bull. London Math. Soc.*, **37**, 243–253.

[21] ——— (2008) Noncommutative Burkholder/Rosenthal Inequalities II: Applications. *Israel J. Math.*, **167**, 227–282.

[22] JUNGE, M. & ZENG, Q. (2011) Noncommutative Bennett and Rosenthal inequalities. Available at `arxiv:1111.1027`.

[23] LEDOUX, M. & TALAGRAND, M. (1991) *Probability in Banach spaces: Isoperimetry and processes*. Springer, Berlin.

[24] LEVINA, E. & VERSHYNIN, R. (2011) Partial estimation of covariance matrices. *Probab. Theory Related Fields*.

[25] LUST-PIQUARD, F. (1986) Inégalités de Khintchine dans C_p $(1 < p < \infty)$. *C. R. Math. Acad. Sci. Paris*, **303**(7), 289–292.

[26] LUST-PIQUARD, F. & PISIER, G. (1991) Noncommutative Khintchine and Paley Inequalities. *Ark. Mat.*, **29**(2), 241–260.

[27] MACKEY, L., JORDAN, M. I., CHEN, R. Y., FARRELL, B. & TROPP, J. A. (2012) Matrix concentration inequalities via the method of exchangeable pairs. Available at `arxiv:1201.6002`.

[28] MARDIA, K. V., KENT, J. T. & BIBBY, J. M. (1980) *Multivariate Analysis*. Academic Press, London.

[29] MUIRHEAD, R. J. (1982) *Aspects of Multivariate Statistical Theory*. Wiley, New York, NY.

[30] OLIVEIRA, R. I. (2010) Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. Available from `arxiv:0911.0600`.

[31] PISIER, G. & XU, Q. (1997) Non-commutative martingale inequalities. *Comm. Math. Phys.*, **189**(3), 667–698.

[32] RECHT, B. (2011) Simpler approach to matrix completion. *J. Mach. Learn. Res.*, **12**, 3413–3430.

[33] ROTHMAN, A. J., LEVINA, E. & ZHU, J. (2009) Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.*, **104**(485), 177–186.

[34] RUDELSON, M. (1999) Random vectors in the isotropic position. *J. Funct. Anal.*, **164**, 60–72.

[35] RUDELSON, M. & VERSHYNIN, R. (2007) Sampling from large matrices: An approach through geometric functional analysis. *J. Assoc. Comput. Mach.*, **54**(4), Article 21, 19 pp., (electronic).

[36] TROPP, J. A. (2008) On the conditioning of random subdictionaries. *Appl. Comput. Harmon. Anal.*, **25**, 1–24.

[37] ——— (2011a) Freedman's inequality for matrix martingales. *Electron. Commun. Probab*, **16**, 262–270.

[38] ——— (2011b) User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*

[39] VERSHYNIN, R. (2011) Introduction to the non-asymptotic analysis of random matrices. in *Compressed Sensing: Theory and Applications*, ed. by Y. Eldar, & G. Kutyniok. Cambridge Univ. Press, Cambridge, Available at `http://www-personal.umich.edu/~romanv/papers/non-asymptotic-rmt-plain.pdf`.